Regression

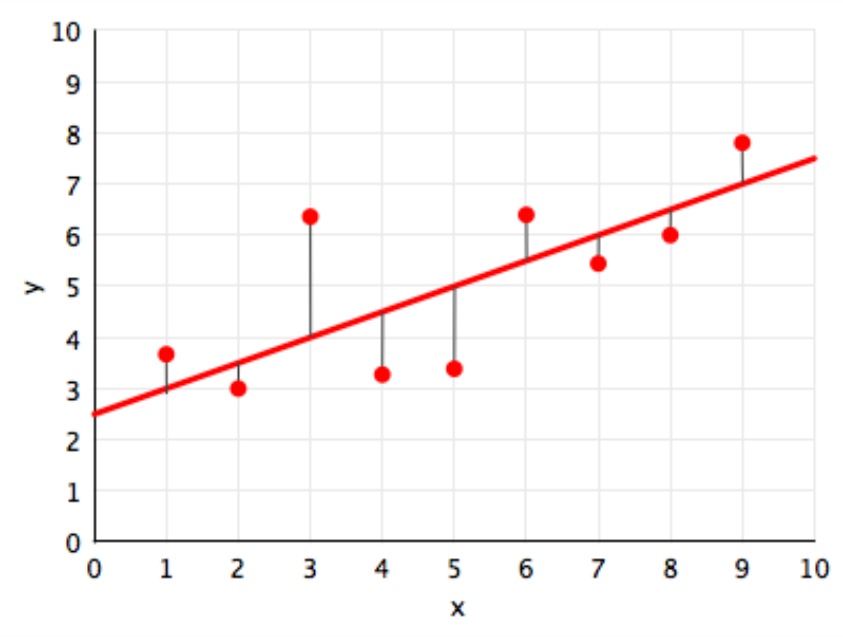
Topics we will cover

1. What is Regression
2. Scoring models
3. Code block A: 1D fit
4. Multi-linear regression
5. Code block B: MLR model fit
6. Code block C: California Real Estate Dataset
7. *Interpreting linear models? (Optional code block)*
8. General paradigm: Model family, loss families

1. What is Regression?

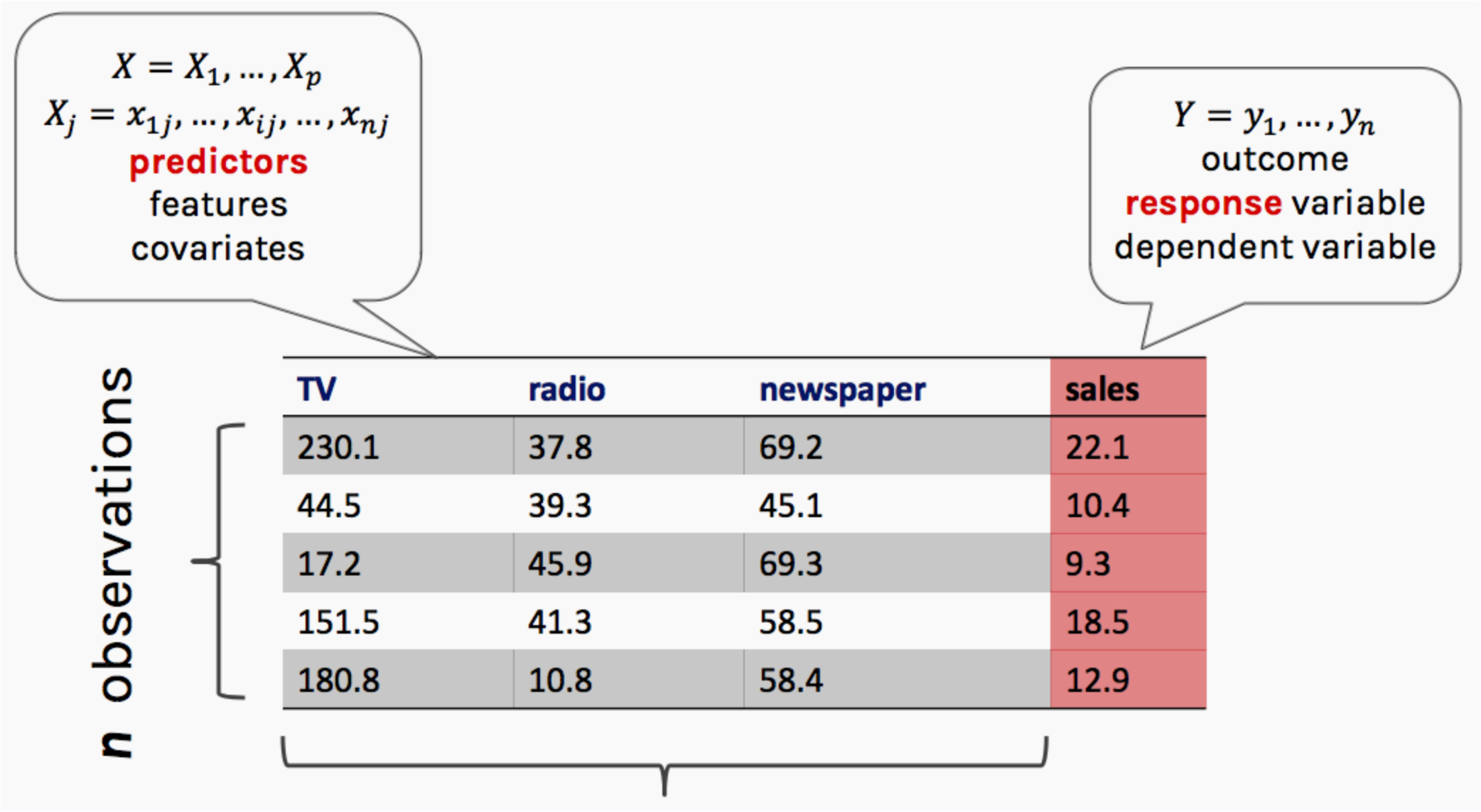
Regression

- how much will you spend?
- what is your creditworthiness
- how many votes t days before election?
- predict probabilities for classification
- risk factors in medicine



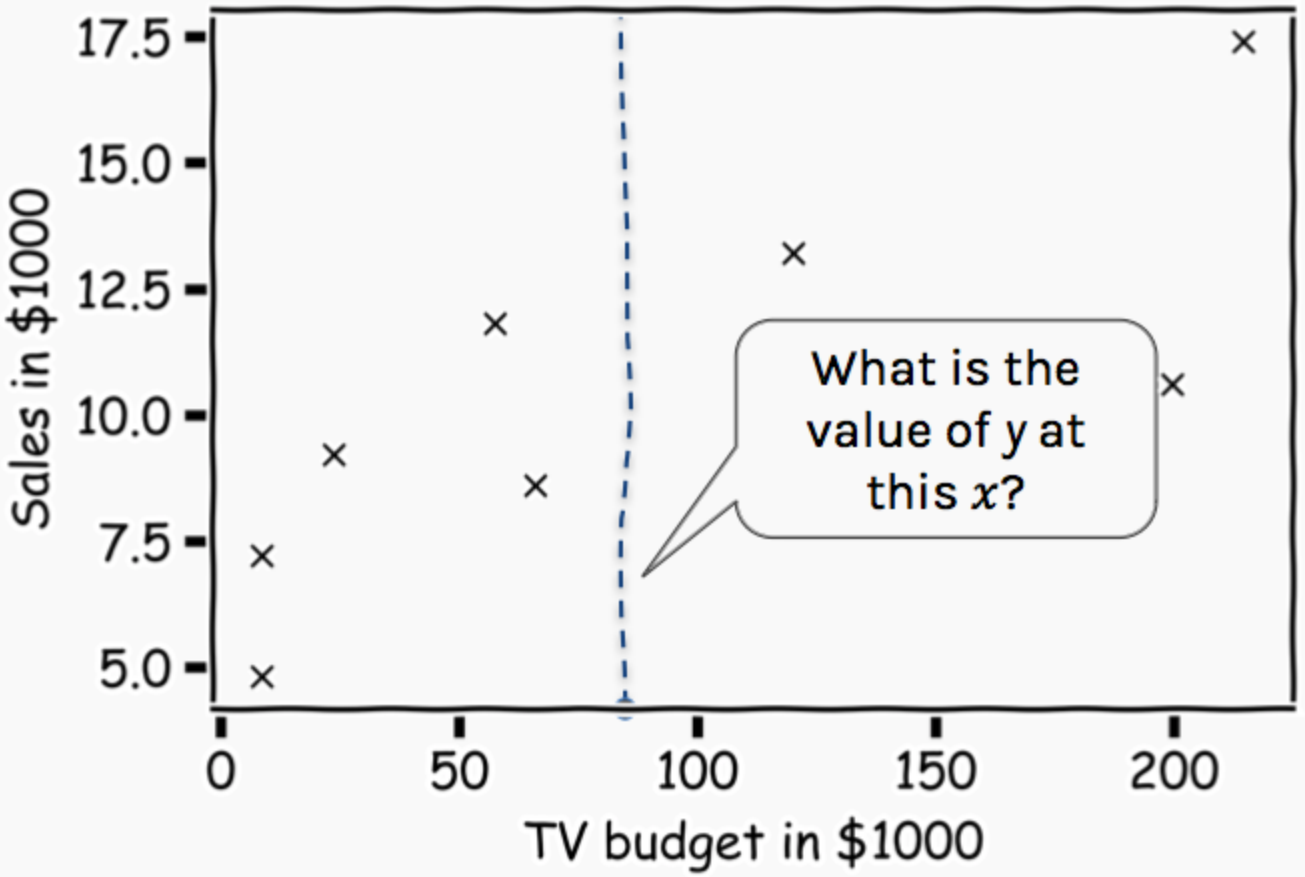
Dataset: Sales and Ad spending

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9


$X = X_1, \ldots, X_p$
$X_j = x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$
predictors
features
covariates
$Y = y_1, \ldots, y_n$
outcome
response variable
dependent variable
n observations
TV
radio
newspaper
sales
230.1
37.8
69.2
22.1
44.5
39.3
45.1
10.4
17.2
45.9
69.3
9.3
151.5
41.3
58.5
18.5
180.8
10.8
58.4
12.9

Two Questions

How do we find $\hat{f}(x)$?



True vs Statistical Model

Truth: Response variable, y, relates to predictors, x, as

$$y = f(x) + \epsilon$$

Here, f **is the unknown function** expressing an underlying rule for relating y to x, and ϵ is random noise (unrelated to x)

In real life we never know the true $f(x)$

Goal: Estimate $f(x)$.

A **statistical model** is any algorithm that estimates f.

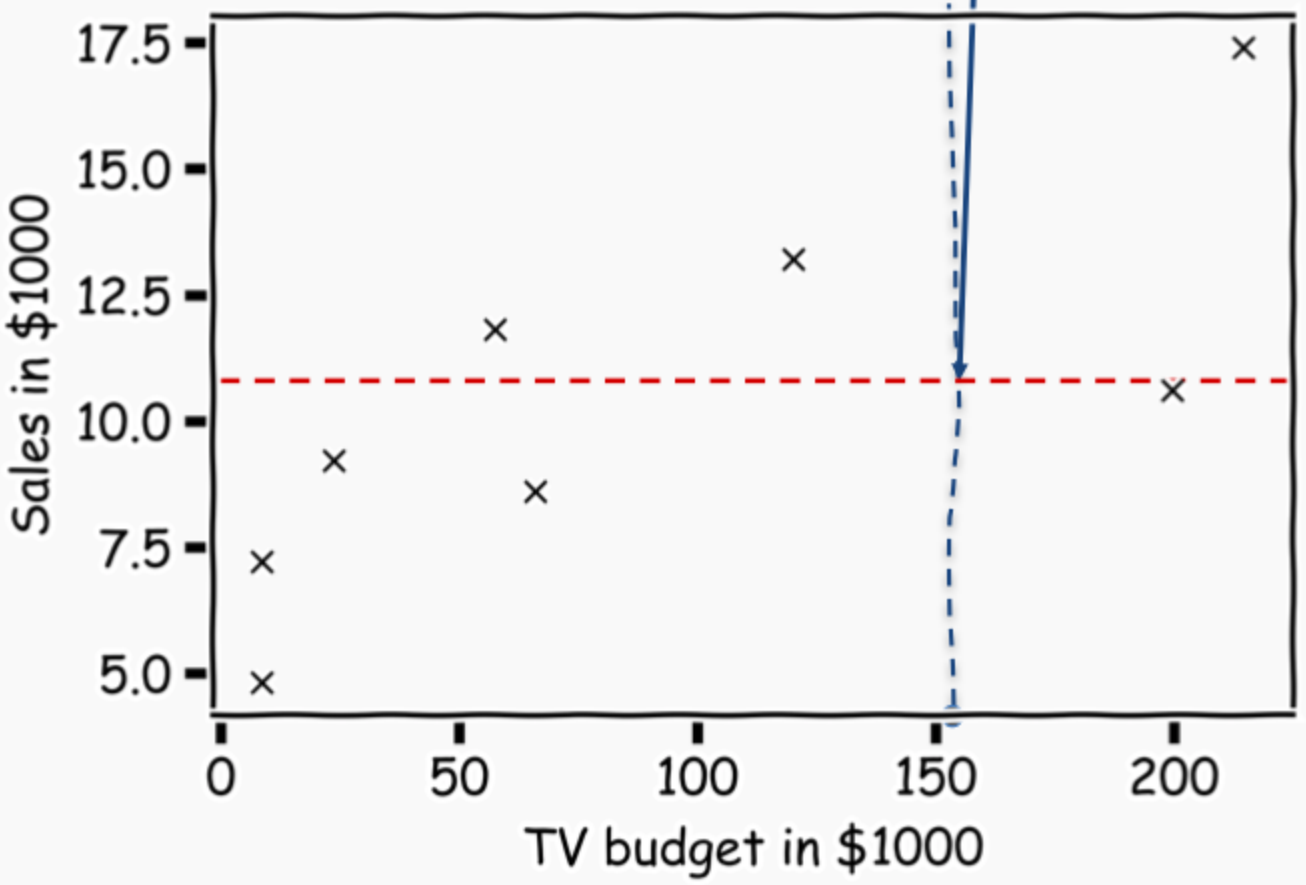
We denote the estimated function as $\hat{f}$.

1. We have no idea about the true generating process. So the function we use may have no relation to the real function that generated the data.
2. We only have access to a sample, with reality denying us access to the entire population.

2. Scoring Models

Simplest model: the mean

Simple idea is to take the mean of all y's, $\hat{f}(x) = \frac{1}{n}\sum_1^n y_i$

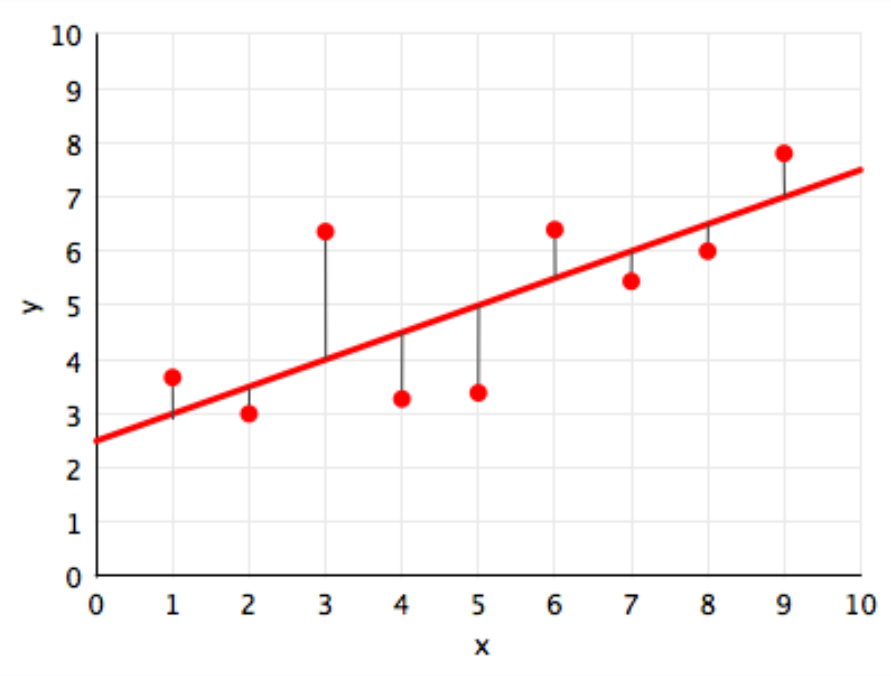


The next simplest: fit a straight line

How? Use **Mean Squared Error**:

$$MSE = \frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2$$

Minimize this with respect to the *parameters*. (intercept and slope)



What kind of loss is this?

For $\hat{y} = a + bx$, the loss is:

$$MSE = \frac{1}{N}\sum_{i}(a + bx_i - y_i)^2$$

This is quadratic and thus convex (bowl shaped).

How can we get to the bottom of the bowl?

Let us assume we can for now ...

What value of the mean squared error is a good one?

The value of the MSE depends on the units of y.

So eliminate dependence on units of y.

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

Evaluation

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

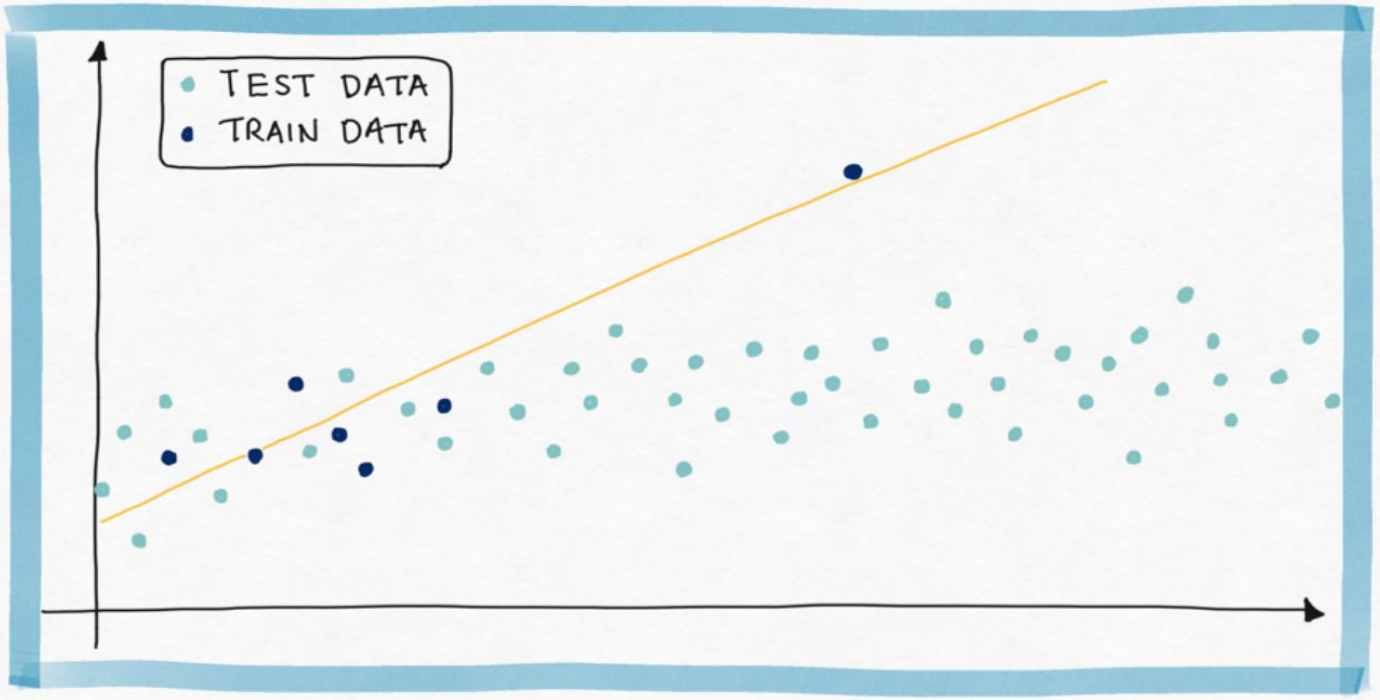
- Model as good as the mean value $\bar{y}$: $R^2 = 0$.
- Model is perfect: $R^2 = 1$.
- R^2 can be negative! You are doing worse than average.
- Can happen when evaluating on the **test set**.

Test vs Train Split

- How do we score models when we don't know ground truth?
- Basic: Hold out some portion of data for training (say 80%) and rest for testing (say 20%)
- **Holiest rule in ML**: Do not contaminate training and test sets!
- More sophisticated: Cross-validation

Evaluation: Test Error

We need to evaluate the fitted model on new data, data that the model did not train on, the test data.



The training MSE here is 2.0 where the test MSE is 12.3.

The training data contains a strange point – an outlier – which confuses the model.

Fitting to meaningless patterns in the training is called **overfitting.**

3. Code block A: Linear Regression in 1D

Question 1:

Which of the following are true about the R^2-score of a model?

1. It can never be negative.
2. It can never be more than 1.
3. Higher is better.
4. The score is 0 when using the mean as our estimator.

4. Multi-linear Regression:

A Case Study

To predict sales vs advertisment would you rather

- Know TV ads budget?
- Know radio ads budget?
- Know newspaper ads budget?
- Know local weather patterns?
- Know all of the above?

Weather seems irrelevant ... if so, we can ingore it!

$X = X_1, \ldots, X_p$

$X_j = x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$

predictors

features

covariates

$Y = y_1, \ldots, y_n$

outcome

response variable

dependent variable

n observations

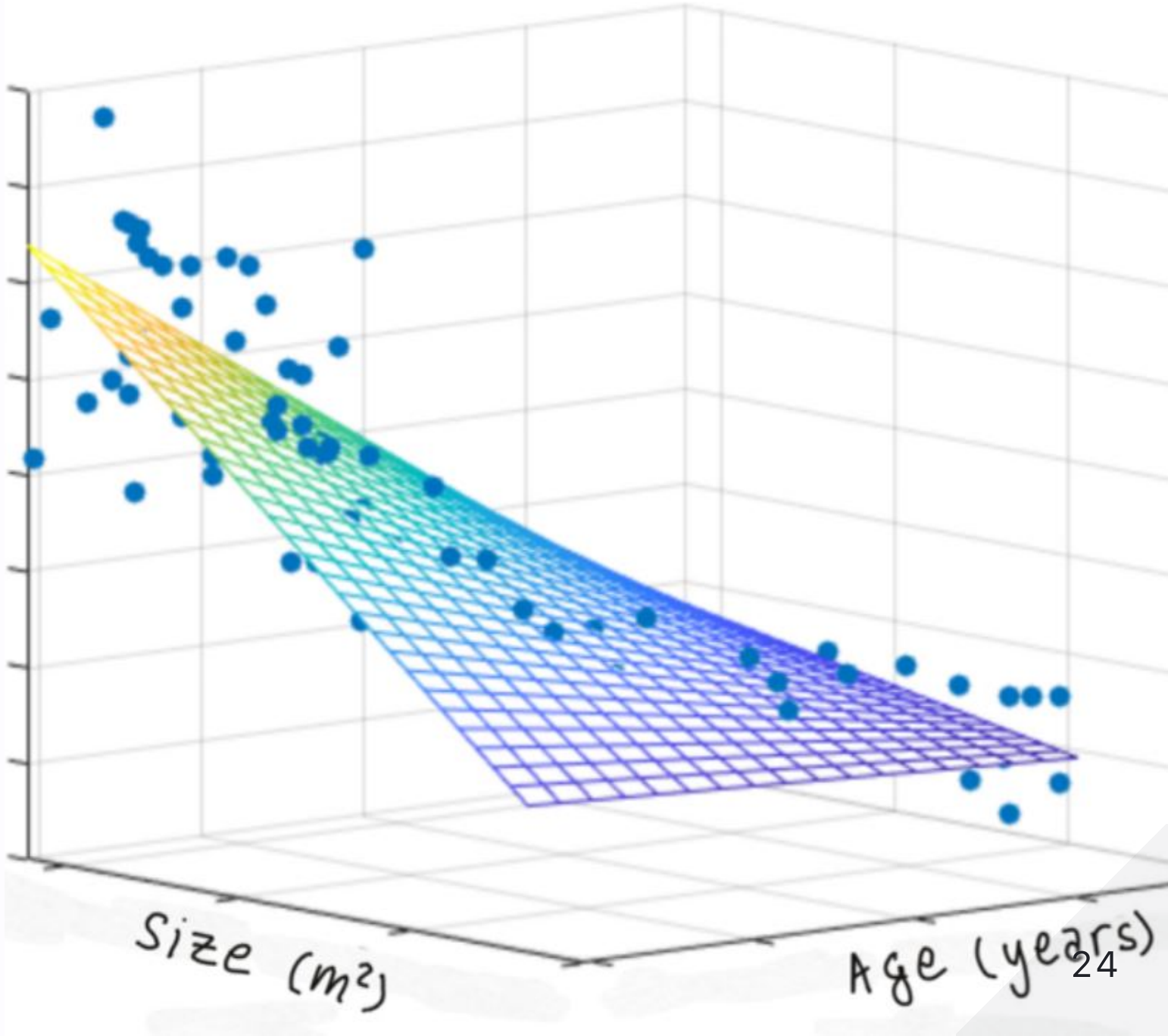
TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Multi-linear Models

- Response variable Y depends on multiple predictors
- Assume predictor variables $x = (x_1, \ldots, x_J)$
- Multi-linear model: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_J x_J$.
- Prediction model $\hat{f}$ has the form:

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_J x_J.$$

Multi-Linear Models



Vector-Notation

Given a set of observations,

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots (x_{n,1}, \ldots, x_{n,J}, y_n)\},$$

the data and the model can be expressed in vector notation,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,J} \\ 1 & x_{2,1} & \ldots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Vector-Notation: Ads Dataset

$$Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Radio + \beta_3 \cdot Newspaper.$$

Linear-algebra notation:

$$\boldsymbol{Y} = \begin{pmatrix} Sales_1 \\ \vdots \\ Sales_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \\ \vdots & \vdots & \vdots \\ 1 & TV_n. & Radio_n & News_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$

Model: $Y = X\beta + \epsilon.$

Multi-Linear Regression: Vector Notation

- Model: $Y = X\beta + \epsilon$.
- Mean-squared error loss: $MSE(\beta) = \frac{1}{n}\|X\beta - Y\|_2^2$.
- There is a closed-form formula for β:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y).$$

(But not too important ... just use solver to fit model)

5. Code block B: Multi-linear regression to predict Sales

6. Code block C: California Housing Dataset

Question 2:

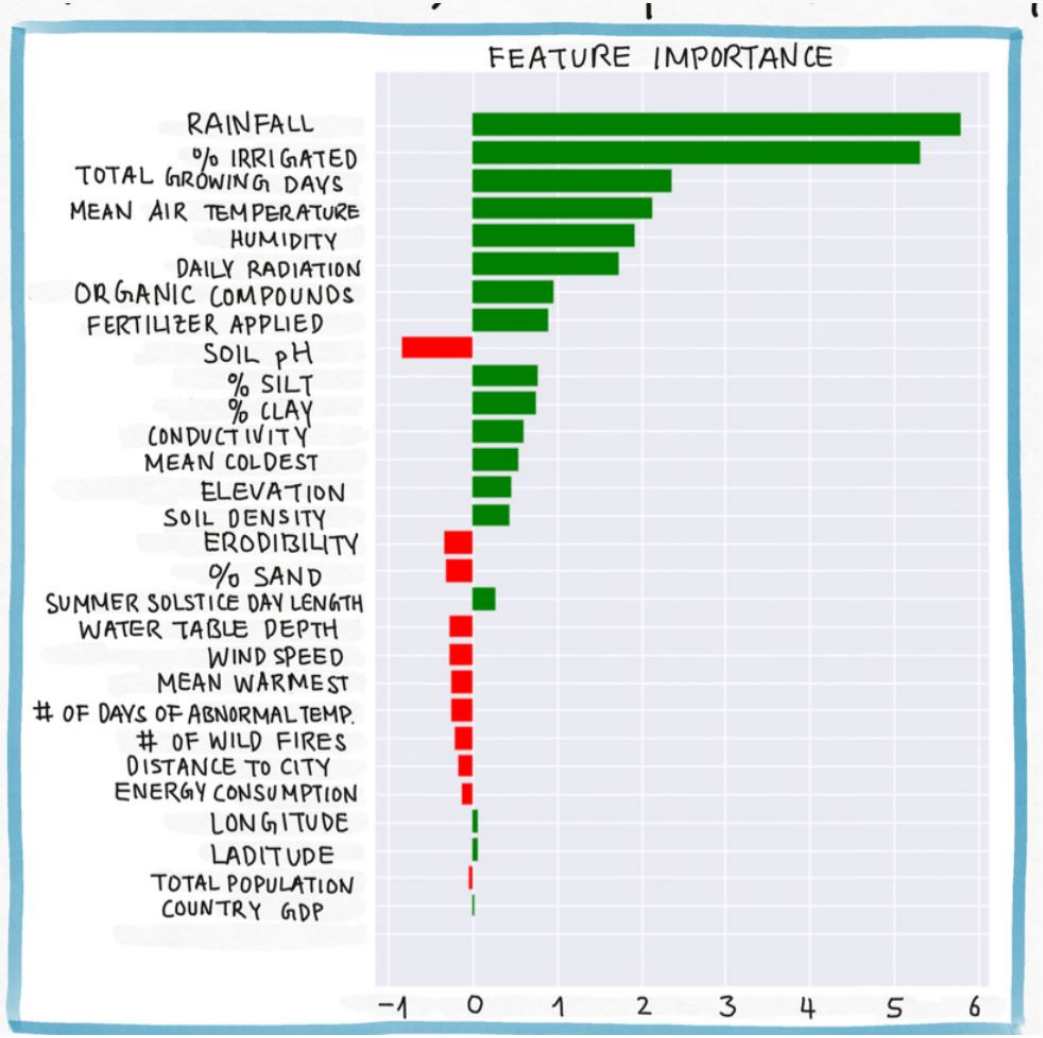
Which of the following statements is the most correct when comparing the performance of a model on the training set versus the performance on the validation set?

1. Performance on the training set will always be better than on the test set.
2. Performance on the training set will typically but not always be better than on the test set.
3. Performance on the trainign set will typically but not always be worse than on the test set.
4. Performance on the trainign set will always be worse than on the test set.

7. Interpretation of Linear Models

Coefficients *can* have meaning

For linear models, it is easy to interpret the model parameters.



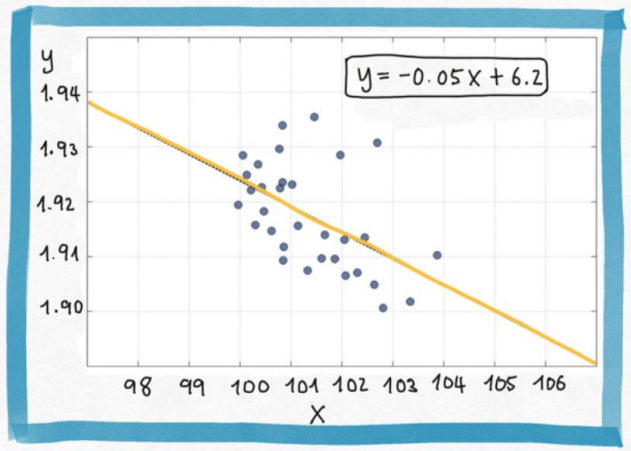
When we have a large number of predictors: $X_1, \ldots, X_J$, there will be a large number of model parameters, $\beta_1, \beta_2, \ldots, \beta_J$.

Looking at the values of β's is impractical, so we visualize these values in a feature importance graph.

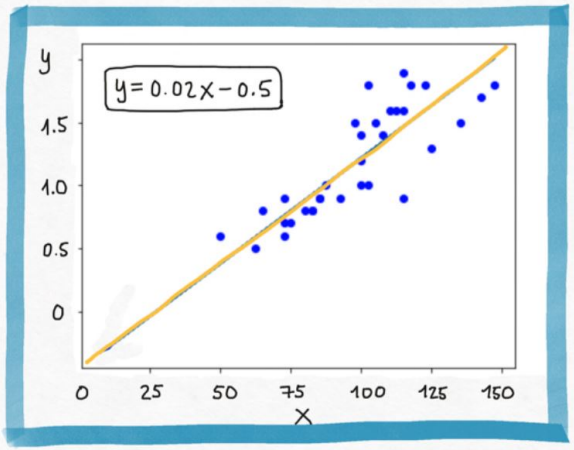
The feature importance graph shows which predictors has the most impact on the model's prediction.

Interpreting Coefficients?

For linear models it's important to interpret the parameters



The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.



The MSE is very small but the intercept is -0.5 which means that for very small budget we will have negative sales.

Are we confident about the interpretation?

Who models the models?

For models like linear regression we can quantify certainty and uncertainty in model parameters.

Bootstrapping Idea: Sub-sample original dataset to build multiple models and take average

Question 3:

Imagine you have a dataset that records if a patient suffered from a heart attack and also lists the following factors for each patient: Age, gender, average blood pressure over six weeks, favorite color.

Which predictors will you consider to make a heart-failure prediction model?

1. Take all predictors
2. Take all except favorite color and gender
3. Take all except favorite color

Question 4:

Imagine you ran a linear regression model for the above dataset and found positive relationship between favorite color being blue and heart-failure.

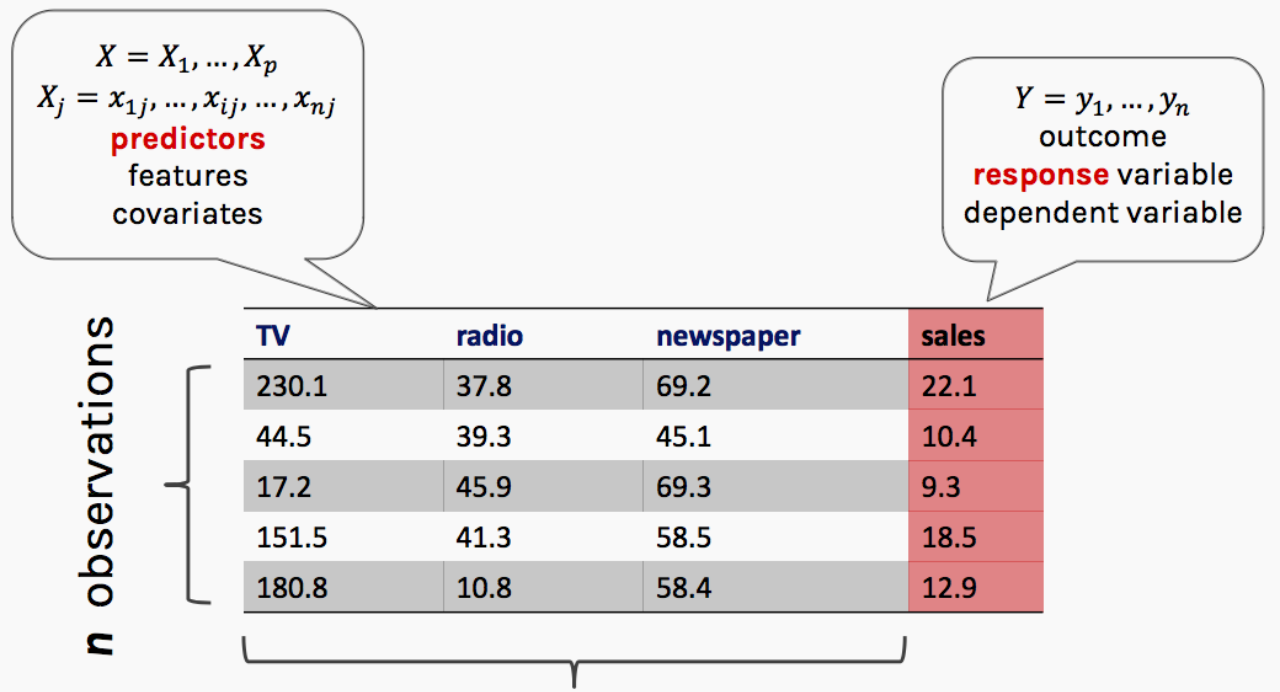
Based on this analysis, the health authorities decide to ban people from wearing blue.

1. This is sound logic.
2. This is absurd. If so, what is a likely explanation?

8. Regression: General Paradigm

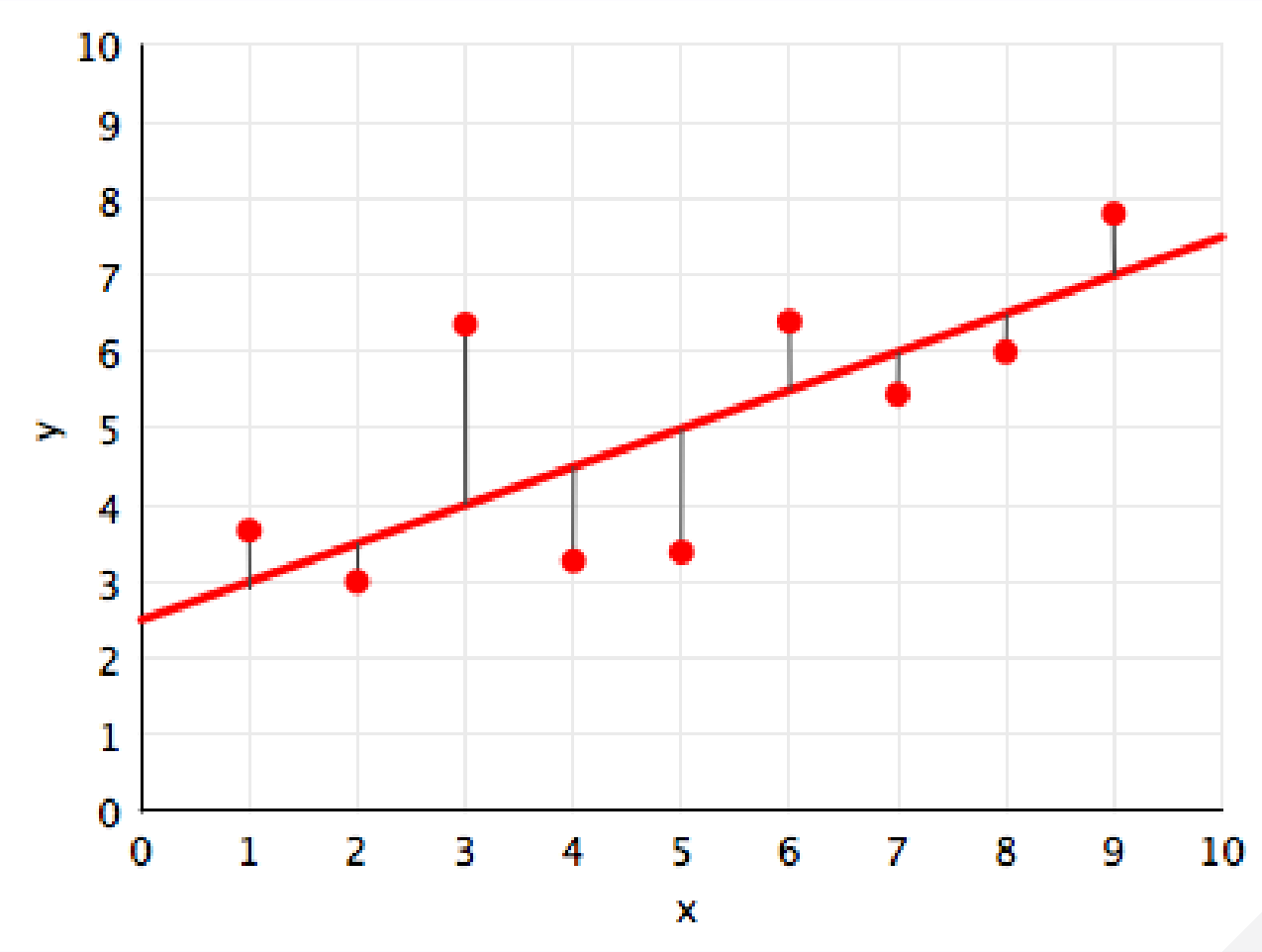
Regression Recipe

- Dataset (X, Y)



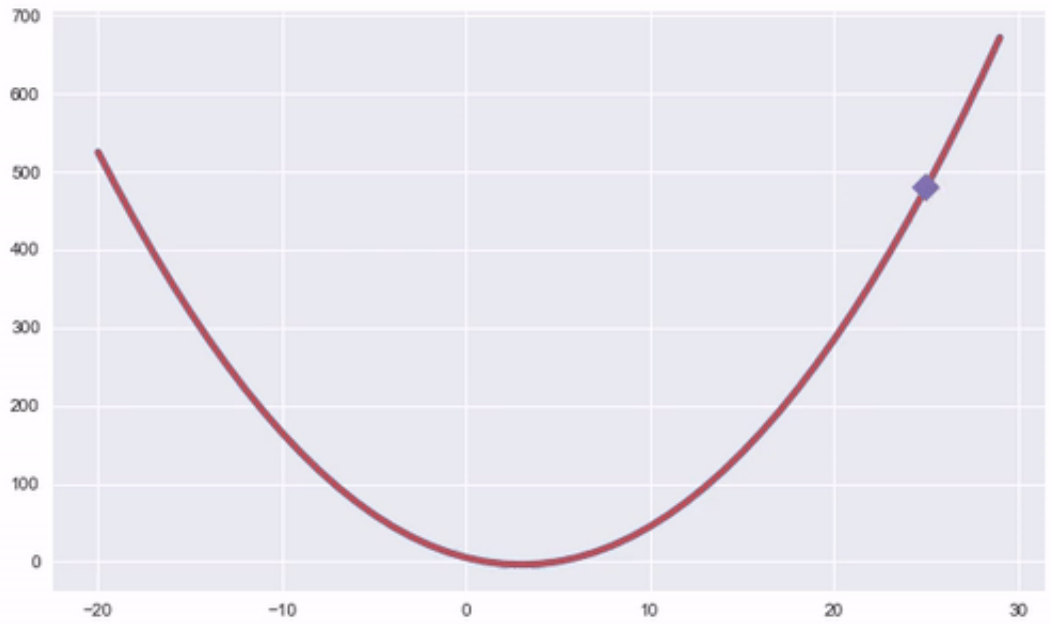
Regression Recipe

- Dataset (X, Y)
- Decide model family: linear, polynomial, ...
- Decide score function: $L(\hat{f}(x), y)$.



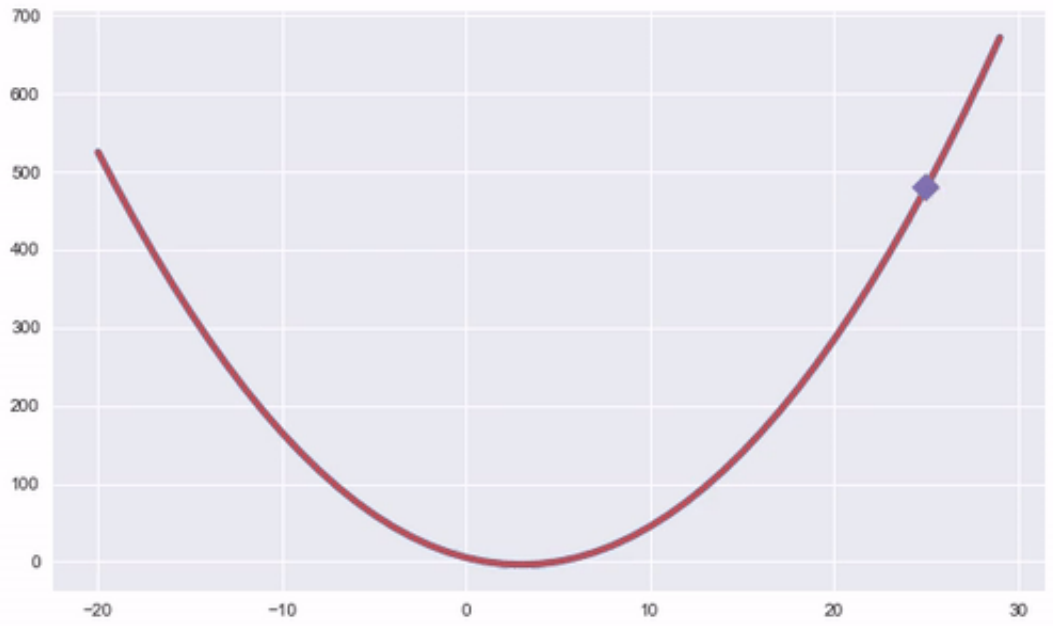
Regression Recipe

- Dataset (X, Y)
- Decide model family: linear, polynomial, ...
- Decide score function: $L(\hat{f}(x), y)$.
- Optimize model
- Use GD, SGD, mini-batch SGD, scikt ...



Future ...

- Dataset (X, Y)
- Decide **model family**: linear, polynomial, ...
- Decide **score function**: $L(\hat{f}(x), y)$.
- Optimize loss
- **Use GD, SGD, mini-batch SGD, scikt** ...



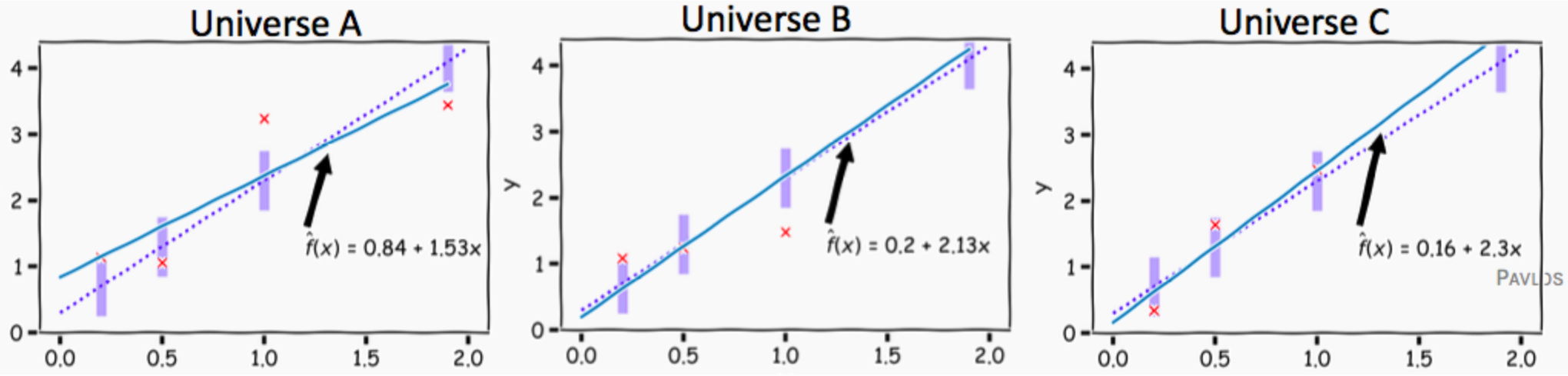
Bonus: Who models the models?

For models like linear regression we can quantify certainty and uncertainty in model parameters.

Bootstrapping Idea: Sub-sample original dataset to build multiple models and take average

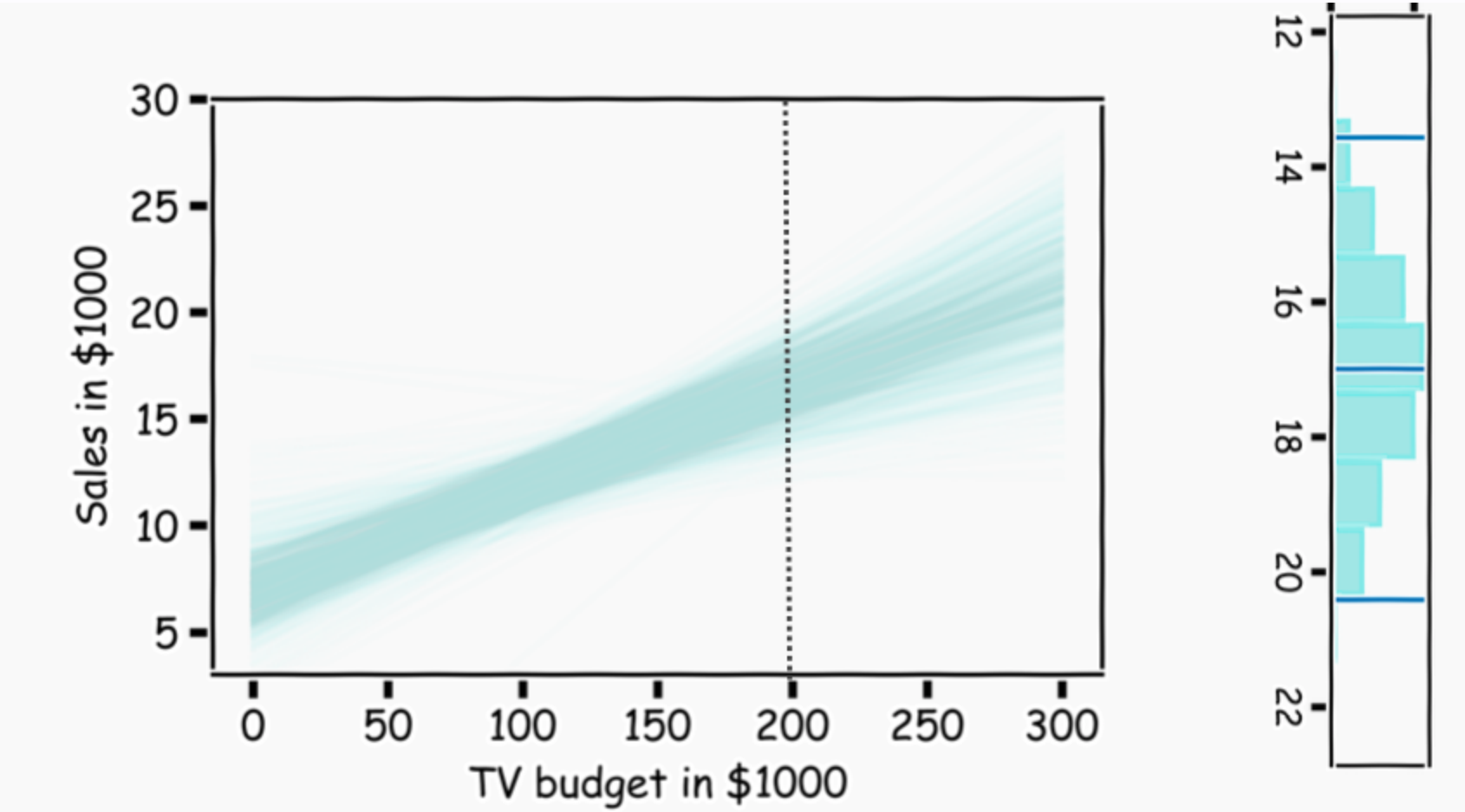
Multiple Fits

- Data is random
- Each dataset gives different model (slope, intercept).

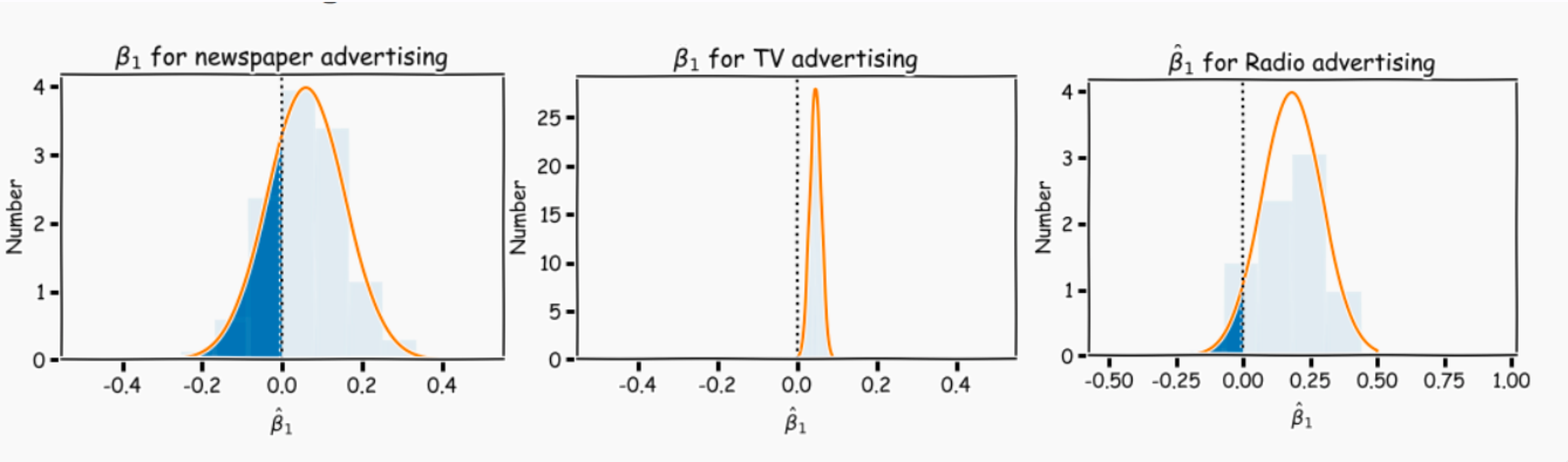


Quantifying Uncertainty in Parameters

- Idea: Plot the different slopes for random subsets
- Compute average and variance of parameters ("standard errors")



Model of models ...



7. Bonus code block D: Model Confidence